The Dynamics of Network Topology

2011 J. Phys.: Conf. Ser. 331 052033

(http://iopscience.iop.org/1742-6596/331/5/052033)

Journal of Physics: Conference Series **331** (2011) 052033 doi:10.1088/1742-6596/331/5/052033

The Dynamics of Network Topology

Ramiro Voicu[1], Iosif Legrand[1], Harvey Newman[1], Artur Barczyk[1], Costin Grigoras[2], Ciprian Dobre[3]

[1] California Institute of Technology, Pasadena, CA 91125, USA

[2] CERN, European Organization for Nuclear Research, CH-1211, Geneve 23, Switzerland

[3] "Politehnica" University of Bucharest, Splaiul Independentei nr. 313, Bucuresti 6, Romania

E-mail: Ramiro.Voicu@cern.ch, Iosif.Legrand@cern.ch, Harvey.Newman@cern.ch, Artur.Barczyk@cern.ch, Costin.Grigoras@cern.ch, Ciprian.Dobre@cs.pub.ro

Abstract. Network monitoring is vital to ensure proper network operation over time, and is tightly integrated with all the data intensive processing tasks used by the LHC experiments. In order to build a coherent set of network management services it is very important to collect in near real-time information about the network topology, the main data flows, traffic volume and the quality of connectivity. A set of dedicated modules were developed in the MonALISA framework to periodically perform network measurements tests between all sites. We developed global services to present in near real-time the entire network topology used by a community. For any LHC experiment such a network topology includes several hundred of routers and tens of Autonomous Systems. Any changes in the global topology are recorded and this information is can be easily correlated with traffic patterns. The evolution in time of global network topology is shown a dedicated GUI. Changes in the global topology at this level occur quite frequently and even small modifications in the connectivity map may significantly affect the network performance. The global topology graphs are correlated with active end to end network performance measurements, done with the Fast Data Transfer application, between all sites. Access to both real-time and historical data, as provided by MonALISA, is also important for developing services able to predict the usage pattern, to aid in efficiently allocating resources globally.

1. Introduction

The monitoring information gathered is essential for developing the required higher level services, the components that provide decision support and some degree of automated decisions and for maintaining and optimizing workflow in large scale distributed systems. Especially the network related aspects as topology monitoring can be very valuable in current LHC era when large amounts of data are expected to be transferred over the network.

1.1. MonALISA Monitoring framework

MonALISA (Monitoring Agents in A Large Integrated Services Architecture) [1] is a globally scalable framework of services developed by Caltech. MonALISA is currently used in several large scale HEP communities and grid systems including CMS [2], ALICE [3], ATLAS [4], the Open Science Grid (OSG) [5], and the Russian LCG sites. It actively monitors USLHCNet [6] production network as well as the UltraLight R&D network [7]. MonALISA also is used to monitor and control all the EVO [8] reflectors, and to help to optimize their interconnections.

As of this writing, more than 300 MonALISA services are running throughout the world. These services monitor more than 60,000 compute servers, and thousands of concurrent jobs. More than 3.5 million parameters are currently monitored in near-real time with an aggregate update rate of approximately 50,000 parameters per second.

A large set of MonALISA monitoring modules has been developed to collect specific network information or to interface it with existing monitoring tools, including:

- SNMP modules for passive traffic measurements and link status
- Active network measurements using simple ping-like measurements
- Tracepath-like measurements to generate the global topology of a wide area network
- Interfaces with the well-known monitoring tools MRTG, RRD [9]
- Available Bandwidth measurements using tools like pathload [10]
- Active bandwidth measurements using Fast Data Transfer (FDT) [11]
- Dedicated modules for TL1 [12] interfaces with CIENA's CD/CIs [13], optical switches (Glimmerglass [14] and Calient [15]) and GMPLS controllers (Calient)

2. Monitoring and representation of network topologies at different OSI layers

We will present in the following subsections briefly both monitoring and representational aspects of network topologies every layer. We had the opportunity inside USLHCNet and UltraLight to have network devices at every OSI layer.

2.1. Physical Network Layer Topology

Specialized TL1 modules are used to monitor optical switches (Layer 1 devices) from two major vendors: Glimmerglass and Calient. We were able to monitor the optical power on ports and the state of the cross-connects inside these switches.

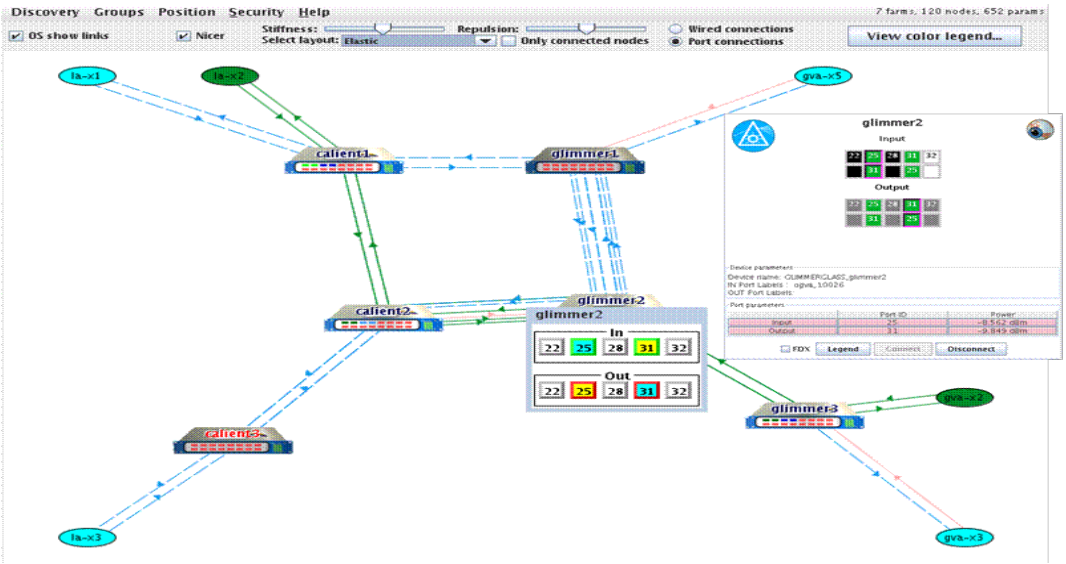

Figure 1 Layer 1 topology: Monitoring and autonomous controlling for optical switches

Based on the monitoring information an agent is able to detect and to take informed decisions in case of eventual problems with the cross connections inside the switch or loss of light on the connections. The MonALISA framework allows one to securely configure many such devices from a single GUI, to see the state of each link in real time, and to have historical plots for the state and activity on each link. It is also easy to manually create a path using the GUI. In Figure 1 we show the MonALISA GUI that is used to monitor the topology on the Layer 1 connections and the state and optical power of the links. The same GUI can be used to request an optical path between any two points in the topology. All the topology related information are kept distributed, every MonALISA service having its own view of the network. Every agent computes a shortest path tree based on Dijkstra's algorithm. The convergence in case of problem is very fast, as every agent has the view of the entire topology.

2.2. Layer 2 Network Topology / Circuits

2.2.1. USLHCNet network

The USLHCNet transatlantic network has evolved from DOE-funded support and management of international networking between the US and CERN. USLHCNet today consists of a backbone of eight 10 Gbps links interconnecting CERN, MANLAN in New York, and Starlight in Chicago. The core of the USLHCNet network is based on Ciena Core Director CD/CI multiplexers which provide stable fallback in case of link outages at Layer 1 (the optical layer), and full support for the GFP/VCAT/LCAS [17] protocol suite.

For the Core Director (CD/CI) we developed modules which monitor the routing protocol (OSRP) which allows us to reconstruct the topology inside the agents, the circuits (VCGs), the state of cross connects, the Ethernet (ETTP/EFLOW) traffic, the allocated time slots on the SONET interfaces and the alarms raised by the CD/CI.

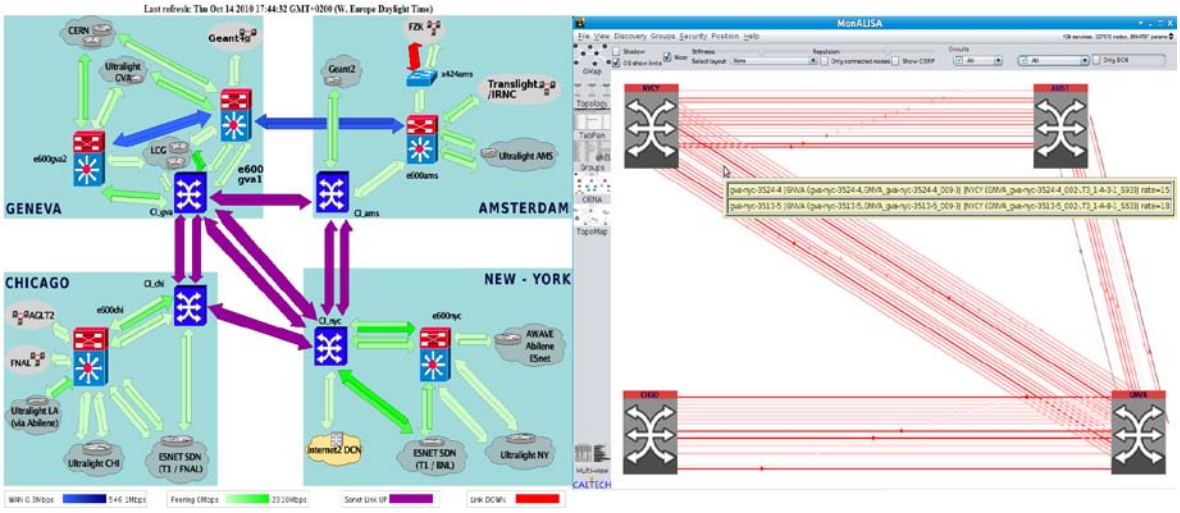

Figure 2 Network weathermap (left) and the layer 2 topology for the dynamic circuits (right)

The operational status for the Force10 ports and all the Ciena CD/CI alarms are recorded by the MonALISA services. They are analyzed and email notification is generated based on different error conditions. We also "monitor" the services used to collect monitoring information. A global repository for all these alarms is available on the MonALISA servers, which allows one to select and sort the alarms based on different conditions. The link status information is very sensitive information for the SLA (Service Level Agreement) with both the experiments and the link providers. Because of this very strict monitoring requirement the monitoring had to have almost 100% availability. This reliability was achieved monitoring each link at both ends from two different points. The NOCs

(Network Operational Center) in Europe, Geneva and Amsterdam, are cross-monitored from both locations, and the same in US. In this way we monitor each link in four points and with special filters this information is directly aggregated in the repository. For redundancy and reliable monitoring we keep at least two instances of repositories running, one in Europe and one in US. For the past two years we manage to have 100% monitoring availability inside USLHCNet.

2.3. Layer 3 Routed Network Topology

For the routed networks, MonALISA is able to construct the overall topology of a complex wide area network, based on the delay on each network segment determined by tracepath-like measurements from each site to all other sites, as it is illustrated in Figure 3.

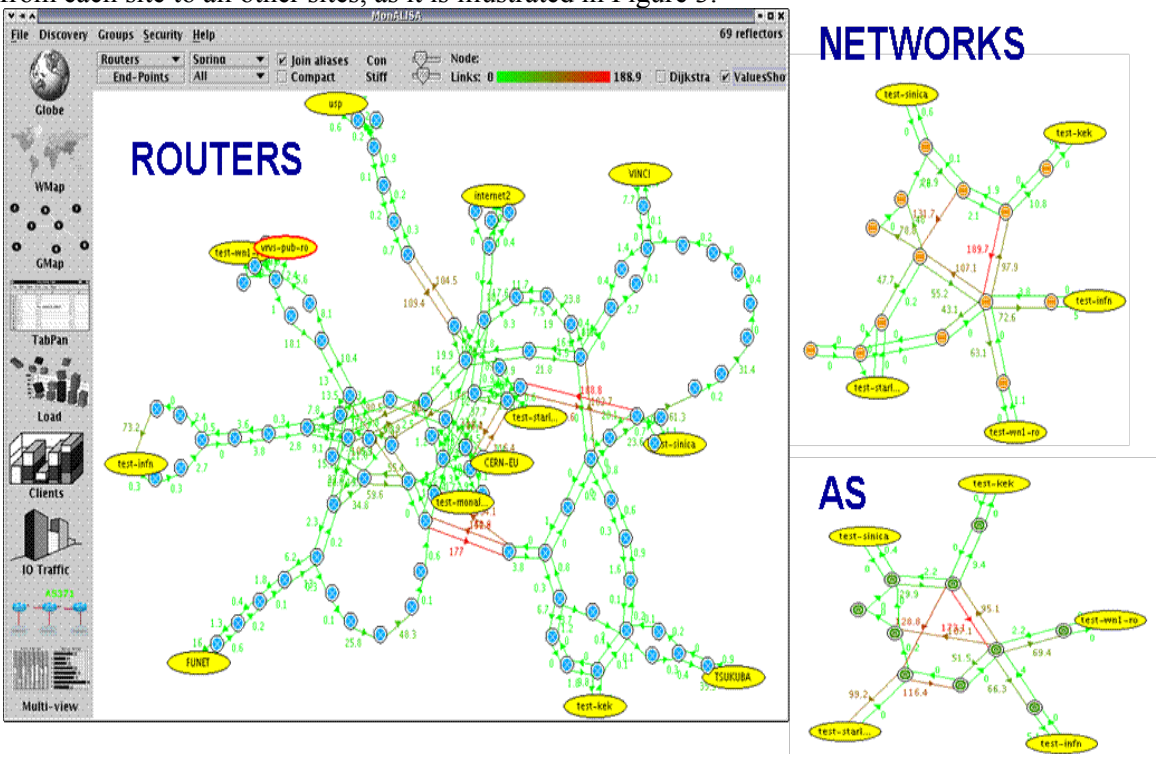

Figure 3 MonALISA real time view of the topology of WANs used by HEP. A view of all the routers, or just the network or "autonomous system" identifiers can be shown.

For any LHC experiment such a network topology includes several hundred of routers and tens of Autonomous Systems. The changes in the global topology are recorded and this information can be easily correlated with traffic patterns. The evolution in time of global network topology is shown a dedicated GUI. Changes in the global topology at this level occur quite frequently and even small modifications in the connectivity map may significantly affect the network performance.

3. A real use case for topology information

The Alice Grid infrastructure uses MonALISA framework for both monitoring and controlling. All the resources used by AliEn [18] services: computing and storage resources, central services, networks, jobs are monitored by MonALISA services at every site.

3.1.1. Bandwidth measurements between Alice sites

The data transfer service is used by the ALICE experiment to perform bandwidth measurements between all sites, by instructing pairs of site MonALISA instances to perform FDT memory-to-memory data transfers with one or more TCP streams.

The results are used for detecting network or configuration problems, since with each test the relevant system configuration and the *tracepath* between the two hosts are recorded as well. The MonALISA services are also used to monitor the end system configuration and automatically notify the user when these systems are not properly configured to support effective data transfers in WAN. In Figure 4 we show the results recorded from one site to all the others.

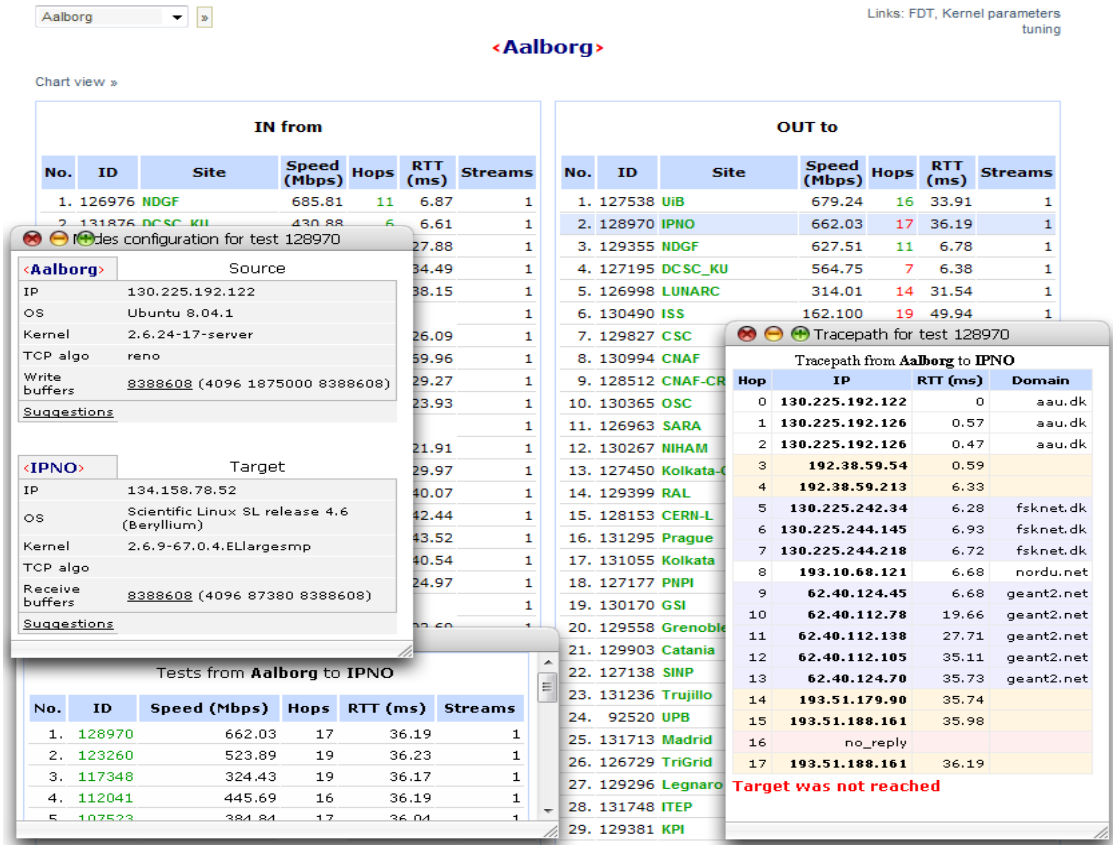

Figure 4 Inter-site bandwidth test results. Tracepath is also recorded

3.1.2. Automatic storage discovery for Alice

Using the monitoring information from trace-like measurements, derived information is computed in the repository, associating the Autonomous System (AS) number to each of the nodes in a network path. The repository also runs other monitoring modules that provide global values and one of them periodically queries AliEn for the list of defined storage elements and their size and usage according to the file catalog. Then periodic functional tests are performed from the central machine to check whether the basic file operations (add, get, remove) are successful. The entire software and network stacks are checked through these tests, thus the outcome should be identical for all clients trying to access the respective storages.

Aggregating the monitoring and test results, a client-to-storage distance metric is computed and used to sort the list of available storage elements to a particular client. Then the closest working storage elements is selected either to save the data or, in case of reading, sorting the available locations based on this metric, trying to read from the closest location. The algorithm associates to each storage element a list of IP addresses representing known machines from its vicinity.

4. Conclusions

We presented in the current paper the capabilities of MonALISA framework towards monitoring and representing network topologies at different OSI layers. We also present a very useful use case where informed automatic decisions based on monitoring information can improve reliability and increase overall performance of the system. In the case of USLHCNet [19], using a distributed monitoring approach, the system achieved 100% monitoring availability, even if the network links are not 100% reliable.

Acknowledgments

This work was supported by the Department of Energy and National Science Foundation within the DoE grant No DE-FG02-08ER41559, by the National Science Foundation within the UltraLight grant, contract No PHY-0427110.

References

[1] MonALISA - MONitoring Agents using a Large Integrated Services Architecture: http://monalisa.caltech.edu
[2] CMS Experiment - The Compact Muon Solenoid Experiment at CERN: http://cms.cern.ch
[3] ALICE Experiment – A Large Ion Collider Experiment at CERN: http://aliweb.cern.ch
[4] Atlas Experiment at CERN: http://atlas.web.cern.ch
[5] OSG – Open Science Grid: http://www.opensciencegrid.org
[6] USLHCNet network: http://uslhcnet.org
[7] UltraLight network: http://ultralight.caltech.edu
[8] EVO – The collaboration network: http://evo.caltech.edu
[9] RRD – The round robin database: http://www.mrtg.org/rrdtool
[10] Pathload – A measurement tool for the available bandwidth of network paths: http://pathload.sourceforge.net
[11] FDT – Fast Data Transfer tool: http://fdt.cern.ch
[12] TL1 – Transaction Language 1 Generic Requirements Document GR-831-CORE: http://telecom-info.telcordia.com/site-cgi/ido/docs.cgi?ID=SEARCH&DOCUMENT=GR-831
[13] Ciena corporation: http://www.ciena.com
[14] Glimmerglass Networks: http://www.glimmerglass.com
[15] Calient Technologies: http://www.calient.net
[16] GMPLS – General Multi-Protocol Label Switching Architecture RFC3945: http://tools.ietf.org/html/rfc3945
[17] ITU-T Rec. G.7042, "Link Capacity Adjustment Scheme (LCAS) for Virtual Concatenated Signals," Feb. 2004.
[18] Bagnasco S, Betev L, Buncic P, Carminati F, Cirstoiu C, Grigoras C, Hayrapetyan A, Harutyunyan A, Peters A J and Saiz P 2007 AliEn : ALICE environment on the GRID, J. Phys.: Conf. Ser. 119
[19] USLHCNet MonALISA Repository: http://repository.uslhcnet.org/